(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 3 506 266 B1

(12) EUROPEAN PATENT SPECIFICATION

(45) Date of publication and mention of the grant of the patent:
09.10.2024 Bulletin 2024/41

(21) Application number: **18209022.5**

(22) Date of filing: **28.11.2018**

(51) International Patent Classification (IPC):
G11C 7/10 (2006.01) ***G11C 11/00*** (2006.01)
G11C 11/54 (2006.01) ***G11C 13/00*** (2006.01)
G06G 7/16 (2006.01) ***G06N 3/063*** (2023.01)
H03M 1/12 (2006.01)

(52) Cooperative Patent Classification (CPC):
G11C 13/0026; G06G 7/16; G11C 7/1006; G11C 11/54; G11C 13/0023; G11C 13/0028; G11C 13/004; G06N 3/065; G11C 2013/0045; G11C 2213/77; H03M 1/12

(54) METHODS AND SYSTEMS FOR PERFORMING A CALCULATION ACROSS A MEMORY ARRAY

VERFAHREN UND SYSTEME ZUR DURCHFÜHRUNG EINER BERECHNUNG ÜBER EINE SPEICHERANORDNUNG HINWEG

PROCÉDÉS ET SYSTÈMES PERMETTANT D'EFFECTUER UN CALCUL SUR UN RÉSEAU DE MÉMOIRE

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR

(30) Priority: **28.12.2017 US 201715857340**

(43) Date of publication of application:
03.07.2019 Bulletin 2019/27

(73) Proprietor: **Tahoe Research, Ltd.**
Dublin 15, D15 YH6H (IE)

(72) Inventors:
- **SRINIVASAN, Srikanth T.**
  Portland, OR 97229 (US)
- **TOMISHIMA, Shigeki**
  Portland, OR 97229 (US)

(74) Representative: **FRKelly**
Waterways House
Grand Canal Quay
Dublin D02 PD39 (IE)

(56) References cited:
WO-A1-2016/099438 **WO-A1-2017/105517**
US-A1- 2016 133 321 **US-A1- 2017 228 345**

- **HARIKA MANEM ET AL: "Design considerations for variation tolerant multilevel CMOS/Nano memristor memory", GREAT LAKES SYMPOSIUM ON VLSI, ACM, 2 PENN PLAZA, SUITE 701 NEW YORK NY 10121-0701 USA, 16 May 2010 (2010-05-16), pages 287 - 292, XP058190497, ISBN: 978-1-4503-0012-4, DOI: 10.1145/1785481.1785548**
- **SUN YULIANG ET AL: "Energy-efficient SQL query exploiting RRAM-based process-in-memory structure", 2017 IEEE 6TH NON-VOLATILE MEMORY SYSTEMS AND APPLICATIONS SYMPOSIUM (NVMSA), IEEE, 16 August 2017 (2017-08-16), pages 1 - 6, XP033163770, DOI: 10.1109/NVMSA.2017.8064463**
- **B. CHAKRABARTI ET AL: "A multiply-add engine with monolithically integrated 3D memristor crossbar/CMOS hybrid circuit", SCIENTIFIC REPORTS, vol. 7, no. 1, 14 February 2017 (2017-02-14), XP055589566, DOI: 10.1038/srep42429**

Processed by Luminess, 75001 PARIS (FR)

Description

BACKGROUND

[0001] Computing device machine learning applications can compute large numbers of Vector-Vector Dot-Products (VVDP). Such large numbers of VVDP computations can incur a corresponding large number of memory accesses to store inputs, store intermediate values, calculate reductions, store reduction values, and the like. The large number of memory accesses associated with computing VVDP can incur substantial memory access delays and/or consume substantially high amounts of power transferring data between processing units and memory, which can create a high computing load on the system.

[0002] US 2016/133321 A1 describes an array of memory cells including resistive memory elements which are coupled to isolation transistors and which include a magnetic tunnel junction. A decoder decodes input address information to select a row of the array. A binarizer coupled to the memory array assigns binary weights to outputs of the memory array output through bit lines coupled to the memory cells. A summer sums the binary weighted outputs, and a quantizer generates an output digital code corresponding to data stored in a plurality of memory cells during a prior program cycle.

[0003] WO 2016/099438 A1 describes a nonvolatile memory cross-bar array including: a number of junctions formed by a number of row lines intersecting a number of column lines; a first set of controls at a first set of the junctions coupling between a first set of the row lines and a first set of the column lines; a second set of controls at a second set of the junctions coupling between a second set of the row lines and a second set of the column lines; and a current collection line to collect currents from the controls of the first set and the second set through their respective column lines and output a result current corresponding to a sum of a first dot product and a second dot product.

[0004] HARIKA MANEM ET AL, "Design considerations for variation tolerant multilevel CMOS/Nano memristor memory", GREAT LAKES SYMPOSIUM ON VLSI, ACM, 2 PENN PLAZA, SUITE 701 NEW YORK NY 10121-0701 USA, (20100516), doi:10.1145/1785481.1785548, ISBN 978-1-4503-0012-4, pages 287 - 292, describes the design of a variation tolerant multilevel CMOS/nano memristor memory.

[0005] SUN YULIANG ET AL, "Energy-efficient SQL query exploiting RRAM-based process-in-memory structure", 2017 IEEE 6TH NON-VOLATILE MEMORY SYSTEMS AND APPLICATIONS SYMPOSIUM (NVMSA), IEEE, (20170816), doi: 10.1109/NVMSA.2017.8064463, pages 1 - 6 describes an RRAM-based SQL query unit with a process in memory structure.

[0006] WO 2017/105517 A1 describes a device to process analog sensor data, the device including at least one analog sensor to generate a first set of analog voltage signals and a crossbar array including a plurality of memristors. The crossbar array is to receive an input vector of the first set of analog voltage signals, generate an output vector comprising a second set of analog voltage signals that is based upon a dot product of the input vector and a matrix comprising resistance values of the plurality of memristors, detect a pattern of the output vector, and activate a processor upon a detection of the pattern.

[0007] US2017/228345 A1 describes a co-processor configured for performing vector matrix multiplication, VMM, to solve computational problems such as partial differential equations, PDEs. An analog Discrete Fourier Transform can be implemented by invoking VMM of input signals with Fourier basis functions using analog crossbar arrays. Linear and non-linear PDEs can be solved by implementing spectral PDE solution methods as an alternative to massively discretized finite difference methods, while exploiting inherent parallelism realized through the crossbar arrays. The analog crossbar array can be implemented in CMOS and memristors or a hybrid solution including a combination of CMOS and memristors.

SUMMARY OF INVENTION

[0008] The present invention is defined in the independent claims 1 and 8.

[0009] Preferred features are recited in the dependent claims. In the following description, any embodiment referred to and not comprising all the features of the independent claims 1 and 8 is merely an example useful to the understanding of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] Features and advantages of the disclosure will be apparent from the detailed description which follows, taken in conjunction with the accompanying drawings, which together illustrate, by way of example, features of the disclosure; and, wherein:

FIG. 1 is a diagram illustrating a computing device programmable media in accordance with an example;
FIG. 2 is a diagram illustrating a computing device readable media in accordance with another example;
FIG. 3 is a diagram illustrating a computing device programmable media in accordance with another example;

FIG. 4 is a diagram illustrating a computing device programmable media in accordance with another example;
FIG. 5 is a diagram illustrating an accelerator in accordance with an example;
FIG. 6 is a diagram illustrating an accelerator in accordance with an example; and
FIG. 7 is a diagram illustrating an accelerator in accordance with an example.

DESCRIPTION OF EMBODIMENTS

[0011] The same reference numerals in different drawings represent the same element. Numbers provided in flow charts and processes are provided for clarity in illustrating steps and operations and do not necessarily indicate a particular order or sequence.

[0012] Furthermore, the described features, structures, or characteristics can be combined in any suitable manner in one or more embodiments. In the following description, numerous specific details are provided, such as examples of layouts, distances, network examples, etc., to convey a thorough understanding of various embodiments. One skilled in the relevant art will recognize, however, that such detailed embodiments do not limit the overall inventive concepts articulated herein, but are merely representative thereof.

[0013] As used in this written description, the singular forms "a," "an" and "the" include express support for plural referents unless the context clearly dictates otherwise. Thus, for example, reference to "a" includes a plurality of such .

[0014] Reference throughout this specification to "an example" means that a particular feature, structure, or characteristic described in connection with the example is included in one or more embodiments. Thus, appearances of the phrases "in an example" or "in an embodiment" in various places throughout this specification are not necessarily all referring to the same embodiment.

[0015] As used herein, a plurality of items, structural elements, compositional elements, and/or materials can be presented in a common list for convenience. However, these lists should be construed as though each member of the list is individually identified as a separate and unique member. Thus, no individual member of such list should be construed as a de facto equivalent of any other member of the same list solely based on their presentation in a common group without indications to the contrary. In addition, various embodiments and examples can be referred to herein along with alternatives for the various components thereof. It is understood that such embodiments, examples, and alternatives are not to be construed as de facto equivalents of one another, but are to be considered as separate and autonomous representations under the present disclosure.

[0016] Furthermore, the described features, structures, or characteristics can be combined in any suitable manner in one or more embodiments. In the following description, numerous specific details are provided, such as examples of layouts, distances, network examples, etc., to provide a thorough understanding of embodiments. One skilled in the relevant art will recognize, however, that the technology can be practiced without one or more of the specific details, or with other methods, components, layouts, etc. In other instances, well-known structures, materials, or operations may not be shown or described in detail to avoid obscuring aspects of the disclosure.

[0017] In this disclosure, "comprises," "comprising," "containing" and "having" and the like can mean "includes," "including," and the like, and are generally interpreted to be open ended terms. The terms "consisting of" or "consists of" are closed terms, and include only the components, structures, steps, or the like specifically listed in conjunction with such terms. In particular, such terms are generally closed terms, with the exception of allowing inclusion of additional items, materials, components, steps, or elements, that do not materially affect the basic and novel characteristics or function of the item(s) used in connection therewith. For example, trace elements present in a composition, but not affecting the composition's nature or characteristics would be permissible if present under the "consisting essentially of" language, even though not expressly recited in a list of items following such terminology. When using an open-ended term in this written description, like "comprising" or "including," it is understood that direct support should be afforded also to "consisting essentially of" language as well as "consisting of" language as if stated explicitly and vice versa.

[0018] The terms "first," "second," "third," "fourth," and the like in the description and in the claims, if any, are used for distinguishing between similar elements and not necessarily for describing a particular sequential or chronological order. It is to be understood that any terms so used are interchangeable under appropriate circumstances such that the embodiments described herein are, for example, capable of operation in sequences other than those illustrated or otherwise described herein. Similarly, if a method is described herein as comprising a series of steps, the order of such steps as presented herein is not necessarily the only order in which such steps may be performed, and certain of the stated steps may possibly be omitted and/or certain other steps not described herein may possibly be added to the method.

[0019] As used herein, comparative terms such as "increased," "decreased," "better," "worse," "higher," "lower," "enhanced," and the like refer to a property of a device, component, or activity that is measurably different from other devices, components, or activities in a surrounding or adjacent area, in a single device or in multiple comparable devices, in a group or class, in multiple groups or classes, or as compared to the known state of the art. For example, a data region that has an "increased" risk of corruption can refer to a region of a memory device, which is more likely to have write errors to it than other regions in the same memory device. A number of factors can cause such increased risk, including location, fabrication process, number of program pulses applied to the region, etc.

[0020] As used herein, the term "substantially" refers to the complete or nearly complete extent or degree of an action, characteristic, property, state, structure, item, or result. For example, an object that is "substantially" enclosed would mean that the object is either completely enclosed or nearly completely enclosed. The exact allowable degree of deviation from absolute completeness may in some cases, depend on the specific context. However, generally speaking, the nearness of completion will be so as to have the same overall result as if absolute and total completion were obtained. The use of "substantially" is equally applicable when used in a negative connotation to refer to the complete or near complete lack of an action, characteristic, property, state, structure, item, or result. For example, a composition that is "substantially free of" particles would either completely lack particles, or so nearly completely lack particles that the effect would be the same as if it completely lacked particles. In other words, a composition that is "substantially free of" an ingredient or element may still actually contain such item as long as there is no measurable effect thereof.

[0021] As used herein, the term "about" is used to provide flexibility to a numerical range endpoint by providing that a given value may be "a little above" or "a little below" the endpoint. However, it is to be understood that even when the term "about" is used in the present specification in connection with a specific numerical value, that support for the exact numerical value recited apart from the "about" terminology is also provided.

[0022] Numerical amounts and data may be expressed or presented herein in a range format. It is to be understood, that such a range format is used merely for convenience and brevity, and thus should be interpreted flexibly to include not only the numerical values explicitly recited as the limits of the range, but also to include all the individual numerical values or sub-ranges encompassed within that range as if each numerical value and sub-range is explicitly recited. As an illustration, a numerical range of "about 1 to about 5" should be interpreted to include not only the explicitly recited values of about 1 to about 5, but also include individual values and sub-ranges within the indicated range. Thus, included in this numerical range are individual values such as 2, 3, and 4 and sub-ranges such as from 1-3, from 2-4, and from 3-5, etc., as well as 1, 1.5, 2, 2.3, 3, 3.8, 4, 4.6, 5, and 5.1 individually.

[0023] This same principle applies to ranges reciting only one numerical value as a minimum or a maximum. Furthermore, such an interpretation should apply regardless of the breadth of the range or the characteristics being described.

[0024] As used herein, the term "circuitry" can refer to, be part of, or include an Application Specific Integrated Circuit (ASIC), an electronic circuit, a processor (shared, dedicated, or group), and/or memory (shared, dedicated, or group) that execute one or more software or firmware programs, a combinational logic circuit, and/or other suitable hardware components that provide the described functionality. In some aspects, the circuitry can be implemented in, or functions associated with the circuitry can be implemented by, one or more software or firmware modules. In some aspects, circuitry can include logic, at least partially operable in hardware.

[0025] Various techniques, or certain aspects or portions thereof, may take the form of program code (i.e., instructions) embodied in tangible media, such as compact disc-read-only memory (CD-ROMs), hard drives, transitory or non-transitory computer readable storage medium, or any other machine-readable storage medium wherein, when the program code is loaded into and executed by a machine, such as a computer, the machine becomes an apparatus for practicing the various techniques. A non-transitory computer readable storage medium can be a computer readable storage medium that does not include signal. In the case of program code execution on programmable computers, the computing device may include a processor, a storage medium readable by the processor (including volatile and non-volatile memory and/or storage elements), an input device, and an output device. The volatile and non-volatile memory and/or storage elements may be a random-access memory (RAM), erasable programmable read only memory (EPROM), flash drive, optical drive, magnetic hard drive, solid state drive, or other medium for storing electronic data. One or more programs that may implement or utilize the various techniques described herein may use an application programming interface (API), reusable controls, and the like. Such programs may be implemented in a high-level procedural or object-oriented programming language to communicate with a computer system. However, the program(s) may be implemented in assembly or machine language, if desired. In any case, the language may be a compiled or interpreted language, and combined with hardware implementations.

[0026] As used herein, the term "processor" can include a single processor or multiple processors, including single core processors and multi-core processors. A processor can include general purpose processors, specialized processors such as central processing units (CPUs), graphics processing units (GPUs), digital signal processors (DSPs), microcontrollers (MCUs), embedded controller (ECs), embedded processors, field programmable gate arrays (FPGAs), network processors, hand-held or mobile processors, application-specific instruction set processors (ASIPs), application-specific integrated circuit processors (ASICs), co-processors, and the like. Additionally, a processor can be packaged in numerous configurations, which is not limiting. For example, a processor can be packaged in a common processor package, a multi-core processor package, a system-on-chip (SoC) package, a system-in-package (SiP) package, a system-on-package (SOP) package, and the like.

[0027] Reference throughout this specification to "an example" or "exemplary" means that a particular feature, structure, or characteristic described in connection with the example is included in one or more embodiments of the present technology. Thus, appearances of the phrases "in an example" or "in an embodiment" or the word "exemplary" in various places throughout this specification are not necessarily all referring to the same embodiment.

Example Embodiments

[0028] An initial overview of technology embodiments is provided below, and then specific technology embodiments are described in further detail later. This initial summary is intended to aid readers in understanding the technology more quickly but is not intended to identify key features or essential features of the technology nor is it intended to limit the scope of the claimed subject matter. The scope of protection of the present invention is defined by appended claims 1-10.

[0029] Vector-Vector Dot-Products (VVDP) are commonly calculated in a variety of applications for a variety of purposes. In one nonlimiting example, VVDP are calculated in machine learning applications to learn useful features for solving problems. In one specific example, a reduction of VVDP can be determined in the analog domain in a computing device programmable media. The computing device programmable media includes a resistive element array, a plurality of bit line select elements and a sense circuit. The resistive element array includes a plurality of word lines, a plurality of bit lines, and a plurality of programmable resistive elements coupled between the plurality of word lines and the plurality of bit lines.

[0030] The resistive elements are programmed to any of a plurality of resistive values. Additionally, the plurality of bit line select elements are configured to couple one or more of the plurality of bit lines to a current summing node in response to one or more of a plurality of bit line sense signals. The sense circuit is configured to sense a parameter of the current summing node in response to a node sense signal.

[0031] The resistive element array is programmed based on values of a weight matrix. Voltages are applied to one or more of the plurality of word lines and/or bit lines based on values of a first vector. One or more of the plurality of bit lines of the resistive element array are selected based on values of a second vector. A reduction value based on the sum of the selected one or more of the plurality of bit lines is thereby determined.

[0032] FIG. 1 is a diagram illustrating a computing device programmable media in accordance with an example. The computing device programmable media may be a separate apparatus, such as an accelerator, or integral to another device or system, such as a Random-Access Memory (RAM). The computing device programmable media includes a resistive element array 105-130 and a reduction circuit 135. The resistive element array includes a plurality of word lines 105-110, a plurality of bit lines 115-120, and a plurality of programmable resistive elements 125-130 coupled between the plurality of word lines 105-110 and the plurality of bit lines 115-120.

[0033] In one implementation, the resistive element array can be a Single Level Cell (SLC) that can store one bit of data, or a Multi-Level Cell (MLC) that can store two or more bits of data. In one implementation, the resistive element array can be any type of byte-accessible memory capable of being used to calculate VVDP. Nonlimiting examples can include phase change memory (PCM), such as chalcogenide glass PCM, planar or 3D PCM, cross-point array memory, including 3D cross-point memory, non-volatile dual in-line memory module (NVDIMM)-based memory, such as persistent memory-based (NVDIMM-P) memory, 3D cross-point-based NVDIMM memory, resistive RAM (ReRAM), including metal-oxide- or oxygen vacancy-based ReRAM, such as HfO_2-, Hf/HfO_x-, Ti/HfO_2-, TiO_x-, and TaO_x-based ReRAM, filament-based ReRAM, such as Ag/GeS_2-, ZrTe/Al_2O_3-, and Ag-based ReRAM, programmable metallization cell (PMC) memory, such as conductive-bridging RAM (CBRAM), silicon-oxide-nitride-oxide-silicon (SONOS) memory, ferroelectric RAM (FeRAM), ferroelectric transistor RAM (Fe-TRAM), anti-ferroelectric memory, polymer memory (e.g., ferroelectric polymer memory), magnetoresistive RAM (MRAM), write-in-place non-volatile MRAM (NVMRAM), spin-transfer torque (STT) memory, spin-orbit torque (SOT) memory, nanowire memory, electrically erasable programmable read-only memory (EEPROM), nanotube RAM (NRAM), other memristor- and thyristor-based memory, spintronic magnetic junction-based memory, magnetic tunneling junction (MTJ)-based memory, domain wall (DW)-based memory, and the like.

[0034] The reduction circuit 135 is configured to select one or more of a plurality of bit lines in response to one or more node select sense signals 140. The reduction circuit 135 is configured to output a reduction value 145 based on a sensed a sum of the selected one or more of the plurality of bit Outputting the reduction value includes saving the reduction value back to the resistive element array.

[0035] The computing device programmable media is utilized to computer reductions of Vector-Vector Dot-Products (VVDP). The resistive element array 105-130 is programmed with resistive values corresponding to an array of weight values. Drive voltages based on a first vector are applied to one or more of the plurality of word lines 105-110 or to one or more of the plurality of bit lines 115-120. One or more node sense signals 140 select one or more of the bit lines 115-120 based on a second vector. A reduction 145 is determined based on a sensed sum of the selected one or more of the plurality of bit lines. Weights of a neural network are loaded into the resistive element array 105-130. Drive voltages corresponding to a first input or intermediate Feature Map (FM) can then be applied to the word lines 105-110 and a second input or intermediate Feature Map (FM) can be applied to the one or more node sense signals 140 to select one or more of the plurality of bit lines 105-110. A reduction 145 is determined based on a sensed sum of the selected one or more of the plurality of bit lines 115-120. Accordingly, VVDPs are performed directly in the computing device programmable media without having to move data between memory and a processing unit of a computing system.

[0036] FIG. 2 is a diagram illustrating a computing device readable media in accordance with another example. Again, the computing device readable media may be a separate apparatus or integral to another device or system. The computing device readable media includes a resistive element array 205-230 and a reduction circuit 235. The resistive element array includes a plurality of word lines 205-210, a plurality of bit lines 215-220 and a plurality of programmable resistive elements 225-230 coupled between the plurality of word lines 205-210 and the plurality of bit lines 215-220.

[0037] In one implementation, the resistive element array can be a Single Level Cell (SLC) that can store one bit of data, or a Multi-Level Cell (MLC) that can store two or more bits of data. In one implementation, the resistive element array can be a resistive cell array, a phase change cell array, a phase change cell array and stackable cross-gridded data access array, a magnetoresistive cell array, a spin torque magnetoresistive cell array, or the like.

[0038] The reduction circuit 235 includes a plurality of bit line select elements 240-245 and a sense circuit 250. The plurality of bit line select elements are configured to couple one or more of the plurality of bit lines 215-220 to a current summing node 255 in response to one or more bit line select signals 260-265. In one implementation, the plurality of bit line select elements 240-245 can include a plurality of Metal Oxide Semiconductor Field Effect Transistors (MOSFET), wherein the gates of the MOSFETs can be coupled to the plurality bit line sense signals 260-265, and the sources and drains of the MOSFETs can be coupled between the plurality of bit lines 215-220 and the current summing node 255. Each respective one of the plurality of MOSFETs can be configured to couple a respective one of the plurality of bit lines 215-220 to the current summing node 255 in response to a respective one of the plurality of bit line select signals 260-265.

[0039] The sense circuit 250 is coupled to the current summing node 255 and configured to sense a parameter of the current summing node 255 in response to the one or more node sense signals 270. The computing device programmable media is utilized to compute reductions of Vector-Vector Dot-Products (VVDP). The resistive element array 205-230 is programmed with resistive values corresponding to an array of weight values. Drive voltages based on a first vector are applied to one or more of the plurality of word lines 205-210, and one or more node bit line select elements 240-245 select one or more of the bit lines 215-220 based on a second vector. A reduction 275 is determined as an output based on a sensed sum by the sense circuit 250 of the selected one or more of the plurality of bit lines 215-220. Weights of a neural network are loaded into the resistive element array 205-230. Drive voltages corresponding to a first input or intermediate Feature Map (FM) can then be applied to the word lines and a second input or intermediate Feature Map (FM) can be applied to the one or more sense signals to select one or more of the plurality of bit lines. For example, a drive voltage V_2 corresponding to a FM value X_2 can be applied to word lines WL2 and a drive voltage V_{127} corresponding to a FM value X_{127} can be applied to word line WL127, and bias voltages can be applied to bit line BL1 selected by select elements SEL1. A reduction is determined based on a sensed sum of the selected one or more of the plurality of bit lines. For example, the current flowing in bit line BL1 can be

$$V_2 \times G_{2-1} + V_{127} \times G_{127-1} = X_2 \times W_{2-1} + X_{127} \times W_{127-1}$$

wherein G_{i-j} is the conductance of the resistive element at row i and column j, and W_{i-j} is the inverse of the conductance. Accordingly, reductions of VVDPs are performed directly in the computing device programmable media without having to move data between memory and a processing unit of a computing system. In one aspect, multiple bit lines in the resistive element array 205-230 can be selected simultaneously to produce multiple VVDPs and by coupling the selected bit lines together, their currents add up to perform the reduction operation. The reductions can be used for normalization of VVDPs.

[0040] FIG. 3 is a diagram illustrating a computing device programmable media in accordance with another example. Again, the computing device programmable media may be a separate apparatus or integral to another device or system. In one implementation, the computing device programmable media can be implemented within a computing device readable memory, such as Random-Access Memory (RAM). In another implementation, the computing device programmable media can be implemented within a computing device accelerator, such as a Vector-Vector Dot-Product (VVDP) reduction accelerator. In another implementation, the computing device programmable media can be implemented as a separate subsystem for use in or by an accelerator, processor, graphics processor or other similar subsystem of a computing system.

[0041] The computing device readable media includes a resistive element array 305-330 and a reduction circuit 335. The resistive element array includes a plurality of word lines 305-310, a plurality of bit lines 315-320 and a plurality of programmable resistive elements 325-330 coupled between the plurality of word lines 305-310 and the plurality of bit lines 315-320.

[0042] In one implementation, the resistive element array can be a Single Level Cell (SLC) that can store one bit of data, or a Multi-Level Cell (MLC) that can store two or more bits of data. In one implementation, the resistive element array can be a resistive cell array, a phase change cell array, a phase change cell array and stackable cross-gridded data access array, a magnetoresistive cell array, a spin torque magnetoresistive cell array, or the like.

[0043] The reduction circuit 335 includes a plurality of bit line select elements 340-345, and a sense circuit 360-370.

In various aspect, the sense circuit 360-370 can be a more detailed implementation of the sense circuit 250 from FIG. 2. The sense circuit 360-370 is configured to output a reduction value 350 based on a sense parameter at a current summing node 355. In one implementation, the sense circuit can include a voltage sense amplifier configured to measure a voltage proportional to the current flowing through the current summing node 355. In another implementation, the sense circuit can include a current sense amplifier configured to measure the current flowing through the current summing node 355. In another implementation, the sense circuit can include an Analog-to-Digital Converter (ADC) 360, a resistive element 365 and a sense gate 370. The resistive element 365 can be coupled between the current summing node 355 and a ground potential. The sense gate 370 can be coupled between the current summing node 355 and the ADC 360. The sense gate 370 can be configured to couple the current summing node 355 to the ADC 360 in response to a node sense signal 375. In one implementation, the sense gate 370 can include a MOSFET including a gate coupled to the node sense signal 375, and a source and drain coupled between the current summing node 355 and the ADC 360. The ADC 360 can be configured to sense an analog voltage value across the resistive element 365 which is proportional to the current flowing through the current summing node 355 and output a digital voltage value as the reduction value 350.

[0044] The computing device programmable media is utilized to computer reductions of Vector-Vector Dot-Products (VVDP). The resistive element array 305-330 is programmed with resistive values corresponding to an array of weight values. Drive voltages based on a first vector are applied to one or more of the plurality of word lines 305-310, and one or more node sense signals 340-345 select one or more of the bit lines 315-320 based on a second vector. A reduction is determined based on a sensed sum of the selected one or more of the plurality of bit lines 315-320. For instance, weights of a neural network are loaded into the resistive element array 305-330. Drive voltages corresponding to a first input or intermediate Feature Map (FM) can then be applied to the word lines 305-310 and a second input or intermediate FM can be applied to the one or more sense signals to select one or more of the plurality of bit lines 315-320. For example, a drive voltage V_2 corresponding to a FM value X_2 can be applied to word lines WL1 and a drive voltage V_{127} corresponding to a FM value X_{127} can be applied to word line WL127, and bias voltages can be applied to bit lines BL1 and BL3 selected by select elements SEL1 SEL3. A reduction is determined based on a sensed sum of the selected one or more of the plurality of bit lines. For example, the current flowing in bit line BL1 can be the sum of I_{1-2} and I_{1-127}, where

$$I_{1-2} = V_2 \times G_{2-1} = X_2 \times W_{2-1}$$

and

$$I_{1-127} = V_{127} \times G_{127-1} = X_{127} \times W_{127-1} ,$$

therefore

$$I_1 = X_2 \times W_{2-1} + X_{127} \times W_{127-1} .$$

Similarly

$$I_3 = X_2 \times W_{2-3} + X_{127} \times W_{127-3} .$$

The current flowing through the sense node 355 can be I_1+I_3., that in turn generates a voltage Vsense = Rdef × (I_1+I_3) across the resistive element 365. The voltage across the resistive element 365 can be sampled by the ADC 360 in response to a node sense signal 375 at the sense gate 370. The summed currents flowing through the sense node 355 represents the reduction of two VVDPs. The selection of even more bit lines 315-320 can be utilized to perform even wider reductions.

Accordingly, reductions of one or more VVDPs can be performed directly in the computing device programmable media without having to move data between memory and a processing unit of a computing system. In one aspect, multiple bit lines in the resistive element array 305-330 can be selected simultaneously to produce multiple VVDPs and by coupling the selected bit lines together, their currents add up to perform the reduction operation. The reductions can be used for normalization of VVDPs.

[0045] FIG. 4 is a diagram illustrating a computing device programmable media in accordance with an example. Again, the computing device programmable media may be a separate apparatus such as an accelerator or integral to another device or system such as a Random-Access Memory (RAM). The computing device readable media can include a resistive element array 405, a word line decoder driver circuit 410, a bit line decoder driver circuit 415, a voltage generator circuit 420, a sense amplifier circuit 425, a page buffer circuit 430, a reduction circuit 435, a reduction buffer circuit 440 and control logic circuit 445.

[0046] The resistive element array 405 can include a plurality of word lines, a plurality of bit lines and a plurality of programmable resistive elements coupled between the plurality of word lines and the plurality of bit lines, as described above with respect to FIGS. 1-3.

[0047] The word line decoder driver circuit 410 and bit line decoder driver circuit 415 can, under control of the control logic 445, decode an address and drive the appropriate word line and bit lines to program a state of resistive element memory cells. The voltage generator circuit 420 can provide appropriate voltage biases to the word line decoder driver circuit 410 and bit line decoder driver circuit 415 for programming cells in the resistive element array 405. For example, the resistive element array 405 can be driven with appropriate voltages to program the resistive elements with values corresponding to an array of weight values, as described above with respect to FIGS. 1-3. Similarly, the word line decoder driver circuit 410 can drive select word lines based on a first vector, the bit line decoder driver circuit 415 can drive the bit lines, and the reduction circuit 435 can select one or more of the bit lines based on a second vector to determine a reduction, as described above with respect to FIGS. 1-3.

[0048] In one or more implementations, the resistive element array 405, the word line decoder driver circuit 410, the bit line decoder driver circuit 415, the voltage generator circuit 420, the sense amplifier circuit 425, the page buffer circuit 430, and the control logic circuit 445 can also operate in accordance with a conventional Resistive Random Access Memory (ReRAM), a Phase Change Memory (PCM), a phase change cell array and stackable cross-gridded data access array (3D XPoint), a Magnetoresistive Random Access Memory (MRAM), a Spin Torque Magnetoresistive Random Access Memory (ST-MRAM), or the like.

[0049] FIG. 5 is a diagram illustrating an accelerator in accordance with an example. The accelerator 505 can include a computing device programmable media 510, a memory controller 515, and a computation controller 520. The accelerator 505 can be communicatively coupled to a host computing device 525. The computing device programmable media 510 can operate to perform reductions of Vector-Vector Dot-Products (VVDP) as described above with respect to FIGS. 1-4.

[0050] In one aspect, the memory controller 515 can receive VVDP requests from the host device 525. The memory controller 515 can be configured to manage the flow of data going to and from the computing device programmable media. The computation controller 520 can be configured to control a reduction circuit of the computing device programmable media 510 to select one or more of a plurality of bit lines and sense a sum of the selected one or more of the plurality of bit lines.

[0051] In one implementation, the accelerator 505 can be separate from the host device 525 as illustrated in FIG. 5. In another implementation, one or more subsystems can be shared between the accelerator 505 and the host device 525, as illustrated in FIG. 6. In such case, one or more portions of the accelerator 505 may be implemented separate from the host device 525, while one or more other portions of the accelerator 505 may be integral to the host device 525. Likewise, one or more portions of the accelerator 505 may be implemented separately or integral to one or more other portions of the accelerator 505. For example, the computation controller 520 can be integral to the memory controller 515 or computing device programmable media 510, or the computation controller 520 can be implemented as a separate portion of the accelerator 505. In addition, one or more sub-systems may be dedicated to implementing the accelerator 505. In other cases, one or more sub-systems may be used in the implementation of the accelerator 505 and used in other sub-systems of the host device 525. For example, the memory controller 515 of the host device 525 may be utilized to read and write data to memory, such as RAM, in the host device 525 and to also manage the flow of data going to and from the computing device programmable media 510 of the accelerator 505. In yet another implementation, the accelerator 505 may be integral to the host device 525 as illustrated in FIG. 7.

[0052] The host computing devices and accelerators of FIGS. 5-7 are illustrative of exemplary embodiments, and are not intended to limit embodiments of the present technology. The accelerator devices and methods described herein can be readily applied to any number of conventional host computing devices, along with computing systems to be developed in the future.

[0053] Embodiments of the present technology are advantageously be utilized in machine learning applications to computer reductions of Vector-Vector Dot-Products (VVDP). Embodiments can be implemented as a dedicated accelerator, or may be implemented as a combination of an accelerator and computer memory. Embodiments advantageously provide a reduction circuit that selects one or more subarrays to produce one or more VVDPs that are summed in the analog domain to perform a reduction operation on the one or more VVDPs. In contrast, the conventional art converts individual VVDP from the analog to digital domain, by means of an analog-to-digital converter, and computing the reduction in the digital domain, by means of digital adders in a processing unit.

[0054] The scope of protection of the present invention is defined in appended claims 1-10.

Claims

1. **1.** An apparatus configured to calculate a reduction of a Vector-Vector Dot-Product (VVDP), comprising:

   a resistive element array (105-130, 205-230, 305-330), including a plurality of word lines (105-110, 205-210,305-310), a plurality of bit lines (115-120, 215-220,315-320), and a plurality of programmable resistive elements (125-130, 225-230, 325-330) coupled between the plurality of word lines and the plurality of bit lines;
   a bit line select circuit comprising a plurality of bit line select elements (240-245, 340-345) configured to couple one or more of the plurality of bit lines to a current summing node (255, 355) in response to one or more of a plurality bit line select signals (140, 260-265, SEL1-SEL128);
   a reduction circuit (135, 235, 335) comprising a sense circuit (250, 360, 370), the sense circuit coupled to the current summing node, wherein the sense circuit is configured to sense a parameter of the current summing node in response to a node sense signal (SEN/SEL, 270, 375); and
   a word line drive circuit and a bit line driver circuit configured to set a state of the plurality of resistive elements according to a first plurality of values based on values of a weight matrix, wherein the word line drive circuit is further configured to drive a selected one or more of the plurality of word lines with a predetermined word line drive parameter according to a second plurality of values,
   wherein the sense circuit is configured to output a reduction value (145, 275, 350) based on the sensed parameter, wherein outputting the reduction value includes saving the reduction value back to the resistive element array.

2. The apparatus of claim 1, wherein the plurality of bit line select elements (240-245) comprises a plurality of Metal Oxide Semiconductor Field Effect Transistors (MOSFET), wherein gates of the MOSFETs are coupled to the plurality bit line select signals (260-265), and sources and drains of the MOSFETs are coupled between the plurality of bit lines (215-220) and the current summing node (250), wherein each respective one of the plurality of MOSFETs is configured to couple a respective one of the plurality of bit lines to the current summing node in response to a respective one of the plurality of bit line select signals.

3. The apparatus of claim 1, wherein the sense circuit includes,

an analog-to-digital converter, ADC, (360);
a resistive element (365) coupled between the current summing node and a ground potential; and
a sense gate (370) coupled between the current summing node and the ADC and configured to couple the current summing node to the ADC in response to the node sense signal, wherein the sense gate comprises a Metal Oxide Semiconductor Field Effect Transistor (MOSFET), wherein the MOSFET includes a gate coupled to the node sense signal and a source and a drain coupled between the current summing node and the ADC.

4. The apparatus of claim 1, wherein,

the plurality of bit line select elements (240-245) couples one or more of the plurality of bit lines to the current summing node in response to the plurality bit line select signals;
the word line drive circuit drives at least two of the plurality of word lines with a predetermined word line drive parameter according to a second plurality of values; and
the sense circuit (250) senses the parameter of the current summing node in response to the node sense signal.

5. The apparatus of claim 1, wherein,

the plurality of bit line select elements (240-245) couples at least two of the plurality of bit lines to the current summing node in response to the plurality of bit line select signals;
the word line drive circuit drives at one or more of the plurality of word lines with a predetermined word line drive parameter according to a second plurality of values; and
the sense circuit (250) senses the parameter of the current summing node in response to the node sense signal.

6. A system, comprising:

the apparatus of claim 1; and
a computation controller (520) configured to control a reduction circuit (235) to select one or more of a plurality of bit lines (215-220) and sense a sum of the selected one or more of the plurality of bit lines; wherein the reduction circuit comprises:

the bit line select circuit (240-245) coupled between the plurality of bit lines and the current summing node, wherein the bit line select circuit is configured to couple the one or more of the bit lines to the current summing node in response to the one or more of a plurality bit line select signals from the computation controller; and
the sense circuit (250) coupled to the current summing node, wherein the sense circuit is configured to sense the parameter of the current summing node in response to the node sense signal from the computation controller and output a reduction value based on the sensed parameter to the memory computation controller.

7. The system of claim 6, wherein the computation controller (520) is further configured to:

control the plurality of word lines and the plurality of bit lines to program the programmable resistive elements according to a first plurality of values;
apply voltages to the plurality of word lines according to a second plurality of values; and
select the one or more of the plurality of bit lines according to a third plurality of values.

8. A method of calculating a reduction of a Vector-Vector Dot-Product (VVDP) comprising:

programming a resistive element array (105-130, 205-230, 305-330) based on values of a weight matrix;
applying voltages to one or more of a plurality of word lines (105-110, 205-210, 305-310) of the resistive element array based on values of a first vector;
coupling a selected one or more of a plurality of bit lines (115-120, 215-220, 315-320) of the resistive element array to a current summing node (255, 355) in response to one or more of a plurality of bit line select signals (140, 260-265, SEL1-SEL128) based on a second vector; sensing a parameter of the current summing node in response to a node sense signal (SEN/ SEL, 270, 375); converting the sensed parameter to a reduction value (145, 275, 350);
outputting the reduction value based on the sensed parameter, and saving the reduction value back to the resistive element array.

9. The method according to claim 8, wherein the weight matrix comprises a Neural Network Matrix (NNM).

10. The method according to claims 8, wherein the first and second vector comprises an input or intermediate Feature Map (FM).

Patentansprüche

1. Einrichtung, ausgelegt zum Berechnen einer Reduktion eines Vektor-Vektor-Skalarprodukts (VVDP), umfassend:

eine Widerstandselementanordnung (105-130, 205-230, 305-330), die eine Vielzahl von Wortleitungen (105-110, 205-210, 305-310), eine Vielzahl von Bitleitungen (115-120, 215-220, 315-320) und eine Vielzahl von programmierbaren Widerstandselementen (125-130, 225-230, 325-330), die zwischen der Vielzahl von Wortleitungen und der Vielzahl von Bitleitungen gekoppelt sind, beinhaltet;
eine Bitleitungsauswahlschaltung, umfassend eine Vielzahl von Bitleitungsauswahlelementen (240-245, 340-345), ausgelegt zum Koppeln einer oder mehrerer der Vielzahl von Bitleitungen mit einem Stromsummierungsknoten (255, 355) als Reaktion auf eines oder mehrere einer Vielzahl von Bitleitungsauswahlsignalen (140, 260-265, SEL1-SEL128);
eine Reduktionsschaltung (135, 235, 335), umfassend eine Erfassungsschaltung (250, 360, 370), wobei die Erfassungsschaltung mit dem Stromsummierungsknoten gekoppelt ist, wobei die Erfassungsschaltung dazu ausgelegt ist, einen Parameter des Stromsummierungsknotens als Reaktion auf ein Knotenerfassungssignal (SEN/SEL, 270, 375) zu erfassen; und
eine Wortleitungsansteuerschaltung und eine Bitleitungsansteuerschaltung, ausgelegt zum Setzen eines Zustands der Vielzahl von Widerstandselementen gemäß ersten mehreren Werten basierend auf Werten einer Gewichtsmatrix, wobei die Wortleitungsansteuerschaltung ferner dazu ausgelegt ist, eine oder mehrere ausgewählte der Vielzahl von Wortleitungen mit einem vorbestimmten Wortleitungsansteuerparameter gemäß einer zweiten Vielzahl von Werten anzusteuern,
wobei die Erfassungsschaltung dazu ausgelegt ist, einen Reduktionswert (145, 275, 350) basierend auf dem erfassten Parameter auszugeben, wobei das Ausgeben des Reduktionswert Speichern des Reduktionswerts

zurück in die Widerstandselementanordnung beinhaltet.

2. Einrichtung nach Anspruch 1, wobei die Vielzahl von Bitleitungsauswahlelementen (240-245) eine Vielzahl von Metall-Oxid-Halbleiter-Feldeffekttransistoren (MOSFET) umfasst, wobei Gates der MOSFETs mit der Vielzahl von Bitleitungsauswahlsignalen (260-265) gekoppelt sind und Sources und Drains der MOSFETs zwischen der Vielzahl von Bitleitungen (215-220) und dem Stromsummierungsknoten (250) gekoppelt sind, wobei jeder jeweilige der Vielzahl von MOSFETs dazu ausgelegt ist, eine jeweilige der Vielzahl von Bitleitungen als Reaktion auf ein jeweiliges der Vielzahl von Bitleitungsauswahlsignalen mit dem Stromsummierungsknoten zu koppeln.

3. Einrichtung nach Anspruch 1, wobei die Erfassungsschaltung Folgendes beinhaltet:

einen Analog-Digital-Wandler, ADC, (360);
ein Widerstandselemente (365), das zwischen dem Stromsummierungsknoten und einem Massepotenzial gekoppelt ist; und
ein Erfassungsgatter (370), das zwischen dem Stromsummierungsknoten und dem ADC gekoppelt und dazu ausgelegt ist, den Stromsummierungsknoten als Reaktion auf das Knotenerfassungssignal mit dem ADC zu koppeln, wobei das Erfassungsgatter einen Metall-Oxid-Halbleiter-Feldeffekttransistoren (MOSFET) umfasst, wobei der MOSFET ein Gate, das mit dem Erfassungsknotensignal gekoppelt ist, und eine Source und einen Drain, die zwischen dem Stromsummierungsknoten und dem ADC gekoppelt sind, beinhaltet.

4. Einrichtung nach Anspruch 1, wobei

die Vielzahl von Bitleitungsauswahlelementen (240-245) eine oder mehrere der Vielzahl von Bitleitungen als Reaktion auf die Vielzahl von Bitleitungsauswahlsignalen mit dem Stromsummierungsknoten koppeln;
die Wortleitungsansteuerschaltung mindestens zwei der Vielzahl von Wortleitungen mit einem vorbestimmten Wortleitungsansteuerparameter gemäß einer zweiten Vielzahl von Werten ansteuert; und
die Erfassungsschaltung (250) den Parameter des Stromsummierungsknotens als Reaktion auf das Knotenerfassungssignal erfasst.

5. Einrichtung nach Anspruch 1, wobei

die Vielzahl von Bitleitungsauswahlelementen (240-245) mindestens zwei der Vielzahl von Bitleitungen als Reaktion auf die Vielzahl von Bitleitungsauswahlsignalen mit dem Stromsummierungsknoten koppelt;
die Wortleitungsansteuerschaltung eine oder mehrere der Vielzahl von Wortleitungen mit einem vorbestimmten Wortleitungsansteuerparameter gemäß einer zweiten Vielzahl von Werten ansteuert; und
die Erfassungsschaltung (250) den Parameter des Stromsummierungsknotens als Reaktion auf das Knotenerfassungssignal erfasst.

6. System, umfassend:

die Einrichtung nach Anspruch 1; und
eine Berechnungssteuerung (520), ausgelegt zum Steuern einer Reduktionsschaltung (235) zum Auswählen einer oder mehrerer einer Vielzahl von Bitleitungen (215-220) und Erfassen einer Summe der einen oder der mehreren ausgewählten der Vielzahl von Bitleitungen; wobei die Reduktionsschaltung Folgendes umfasst:

die Bitleitungsauswahlschaltung (240-245), gekoppelt zwischen der Vielzahl von Bitleitungen und dem Stromsummierungsknoten, wobei die Bitleitung aus Wandlerschaltung ausgelegt ist zum Koppeln der einen oder der mehreren der Bitleitungen mit dem Stromsummierungsknoten als Reaktion auf das eine oder die mehreren einer Vielzahl von Bitleitungsauswahlsignalen für die Berechnungssteuerung; und
die Erfassungsschaltung (250), gekoppelt mit dem Stromsummierungsknoten, wobei die Erfassungsschaltung ausgelegt ist zum Erfassen des Parameters des Stromsummierungsknoten als Reaktion auf das Knotenerfassungssignal von der Berechnungssteuerung und Ausgeben eines Reduktionswerts basierend auf dem erfassten Parameter an die Speicherberechnungssteuerung.

7. System nach Anspruch 6, wobei die Berechnungssteuerung (520) ferner zu Folgendem ausgelegt ist:

Steuern der Vielzahl von Wortleitungen und der Vielzahl von Bitleitungen zum Programmieren der programmierbaren Widerstandselemente gemäß einer ersten Vielzahl von Werten;

Anlegen von Spannungen an die Vielzahl von Wortleitungen gemäß einer zweiten Vielzahl von Werten; und Auswählen der einen oder der mehreren der Vielzahl von Bitleitungen gemäß einer dritten Vielzahl von Werten.

8. Verfahren zum Berechnen einer Reduktion eines Vektor-Vektor-Skalarprodukts (VVDP), umfassend:

Programmieren einer Widerstandselementanordnung (105-130, 205-230, 305-330) basierend auf Werten einer Gewichtsmatrix;
Anlegen einer Spannung an eine oder mehrere einer Vielzahl von Wortleitungen (105-110, 205-210, 305-310) der Widerstandselementanordnung basierend auf Werten eines ersten Vektors;
Koppeln einer oder mehrerer ausgewählter einer Vielzahl von Bitleitungen (115-120, 215-220, 315-320) der Widerstandselementanordnung mit einem Stromsummierungsknoten (255, 355) als Reaktion auf ein oder mehrere einer Vielzahl von Bitleitungsauswahlsignalen (140, 260-265, SEL1-SEL128) basierend auf einem zweiten Vektor; Erfassen eines Parameters des Stromsummierungsknotens als Reaktion auf ein Knotenerfassungssignal (SEN/SEL, 270, 375);
Umwandeln des erfassten Parameters in einen Reduktionswert (145, 275, 350); Ausgeben des Reduktionswerts basierend auf dem erfassten Parameter und Speichern des Reduktionswerts zurück in die Widerstandselementanordnung.

9. Verfahren nach Anspruch 8, wobei die Gesichtsmatrix eine Neural Network Matrix (NNM) umfasst.

10. Verfahren nach Anspruch 8, wobei der erste und der zweite Vektor eine Eingabe- oder Zwischen-Feature-Map (FM) umfassen.

Revendications

1. Appareil configuré pour calculer une réduction d'un produit scalaire vecteur-vecteur (VVDP), comprenant :

un ensemble d'éléments résistifs (105-130, 205-230, 305-330), comportant une pluralité de lignes de mots (105-110, 205-210, 305-310), une pluralité de lignes de bits (115-120, 215-220, 315-320) et une pluralité d'éléments résistifs programmables (125-130, 225-230, 325-330) couplés entre la pluralité de lignes de mots et la pluralité de lignes de bits ;
un circuit de sélection de ligne de bits comprenant une pluralité d'éléments de sélection (240-245, 340-345) de ligne de bits configurés pour coupler une ou plusieurs de la pluralité de lignes de bits à un noeud de sommation de courant (255, 355) en réponse à un ou plusieurs d'une pluralité de signaux de sélection de ligne de bits (140, 260-265, SEL1-SEL128) ;
un circuit de réduction (135, 235, 335) comprenant un circuit de détection (250, 360, 370), le circuit de détection étant couplé au noeud de sommation de courant, le circuit de détection étant configuré pour détecter un paramètre du noeud de sommation de courant en réponse à un signal de détection de noeud (SEN/SEL, 270, 375) ;
et un circuit de commande de ligne de mots et un circuit de commande de ligne de bits configurés pour définir un état de la pluralité d'éléments résistifs en fonction d'une première pluralité de valeurs sur la base de valeurs d'une matrice de poids, le circuit de commande de ligne de mots étant en outre configuré pour commander une ou plusieurs lignes sélectionnées de la pluralité de lignes de mots avec un paramètre de commande de ligne de mots prédéterminé en fonction d'une deuxième pluralité de valeurs,
le circuit de détection étant configuré pour délivrer une valeur de réduction (145, 275, 350) basée sur le paramètre détecté, la fourniture de la valeur de réduction comportant la sauvegarde de la valeur de réduction de nouveau vers l'ensemble d'éléments résistifs.

2. Appareil selon la revendication 1, la pluralité d'éléments de sélection (240-245) de ligne de bits comprenant une pluralité de transistors à effet de champ à métal-oxyde-semiconducteurs (MOSFET), des grilles des MOSFET étant couplées à la pluralités de signaux de sélection de ligne de bits (260-265) et des sources et des drains des MOSFET étant couplés entre la pluralité de lignes de bits (215-220) et le noeud de sommation de courant (250), chaque MOSFET respectif de la pluralité de MOSFET étant configuré pour coupler une ligne respective de la pluralité de lignes de bits au noeud de sommation de courant en réponse à un signal respectif de la pluralité de signaux de sélection de ligne de bits.

3. Appareil selon la revendication 1, le circuit de détection comportant,

un convertisseur analogique-numérique, CAN, (360) ;
un élément résistif (365) couplé entre le noeud de sommation de courant et un potentiel de masse ; et
une grille de détection (370) couplée entre le noeud de sommation de courant et le CAN et configurée pour coupler le noeud de sommation de courant au CAN en réponse au signal de détection de noeud, la grille de détection comprenant un transistor à effet de champ à métal-oxyde-semiconducteur (MOSFET), le MOSFET comportant une grille couplée au signal de détection de noeud et une source et un drain couplés entre le noeud de sommation de courant et le CAN.

4. Appareil selon la revendication 1,

la pluralité d'éléments de sélection (240-245) de ligne de bits couplant une ou plusieurs de la pluralité de lignes de bits au noeud de sommation de courant en réponse à la pluralité de signaux de sélection de ligne de bits ;
le circuit de commande de ligne de mots commandant au moins deux de la pluralité de lignes de mots avec un paramètre de commande de ligne de mots prédéterminé en fonction d'une deuxième pluralité de valeurs ; et
le circuit de détection (250) détectant le paramètre du noeud de sommation de courant en réponse au signal de détection de noeud.

5. Appareil selon la revendication 1,

la pluralité d'éléments de sélection (240-245) de ligne de bits couplant au moins deux de la pluralité de lignes de bits au noeud de sommation de courant en réponse à la pluralité de signaux de sélection de ligne de bits ;
le circuit de commande de ligne de mots commandant une ou plusieurs de la pluralité de lignes de mots avec un paramètre de commande de ligne de mots prédéterminé en fonction d'une deuxième pluralité de valeurs ; et
le circuit de détection (250) détectant le paramètre du noeud de sommation de courant en réponse au signal de détection de noeud.

6. Système, comprenant :

l'appareil selon la revendication 1 ; et
un contrôleur de calcul (520) configuré pour commander un circuit de réduction (235) pour sélectionner une ou plusieurs d'une pluralité de lignes de bits (215-220) et détecter une somme des une ou plusieurs lignes sélectionnées de la pluralité de lignes de bits ; le circuit de réduction comprenant :

le circuit de sélection (240-245) de ligne de bits couplé entre la pluralité de lignes de bits et le noeud de sommation de courant, le circuit de sélection de ligne de bits étant configuré pour coupler les une ou plusieurs des lignes de bits au noeud de sommation de courant en réponse aux un ou plusieurs d'une pluralité de signaux de sélection de ligne de bits provenant du contrôleur de calcul ; et
le circuit de détection (250) couplé au noeud de sommation de courant, le circuit de détection étant configuré pour détecter le paramètre du noeud de sommation de courant en réponse au signal de détection de noeud provenant du contrôleur de calcul et délivrer une valeur de réduction basée sur le paramètre détecté vers le contrôleur de calcul de mémoire.

7. Système selon la revendication 6, le contrôleur de calcul (520) étant en outre configuré pour :

commander la pluralité de lignes de mots et la pluralité de lignes de bits pour programmer les éléments résistifs programmables en fonction d'une première pluralité de valeurs ;
appliquer des tensions à la pluralité de lignes de mots en fonction d'une deuxième pluralité de valeurs ; et
sélectionner les une ou plusieurs de la pluralité de lignes de bits en fonction d'une troisième pluralité de valeurs.

8. Procédé de calcul d'une réduction d'un produit scalaire vecteur-vecteur (VVDP) comprenant :

la programmation d'un ensemble d'éléments résistifs (105-130, 205-230, 305-330) sur la base de valeurs d'une matrice de poids ;
l'application de tensions à une ou plusieurs d'une pluralité de lignes de mots (105-110, 205-210, 305-310) de l'ensemble d'éléments résistifs sur la base de valeurs d'un premier vecteur ;
le couplage d'une ou plusieurs lignes sélectionnées d'une pluralité de lignes de bits (115-120, 215-220, 315-320) de l'ensemble d'éléments résistifs à un noeud de sommation de courant (255, 355) en réponse à un ou plusieurs d'une pluralité de signaux de sélection de ligne de bits (140, 260-265, SEL1-SEL128) sur la base d'un deuxième

vecteur ;
la détection d'un paramètre du noeud de sommation de courant en réponse à un signal de détection de noeud (SEN/SEL, 270, 375) ; la conversion du paramètre détecté en une valeur de réduction (145, 275, 350) ; la fourniture de la valeur de réduction basée sur le paramètre détecté et la sauvegarde de la valeur de réduction de nouveau vers l'ensemble d'éléments résistifs.

9. Procédé selon la revendication 8, la matrice de poids comprenant une matrice de réseau de neurones artificiels (NNM).

10. Procédé selon la revendication 8, les premier et deuxième vecteurs comprenant une carte de caractéristiques (FM) d'entrée ou intermédiaire.

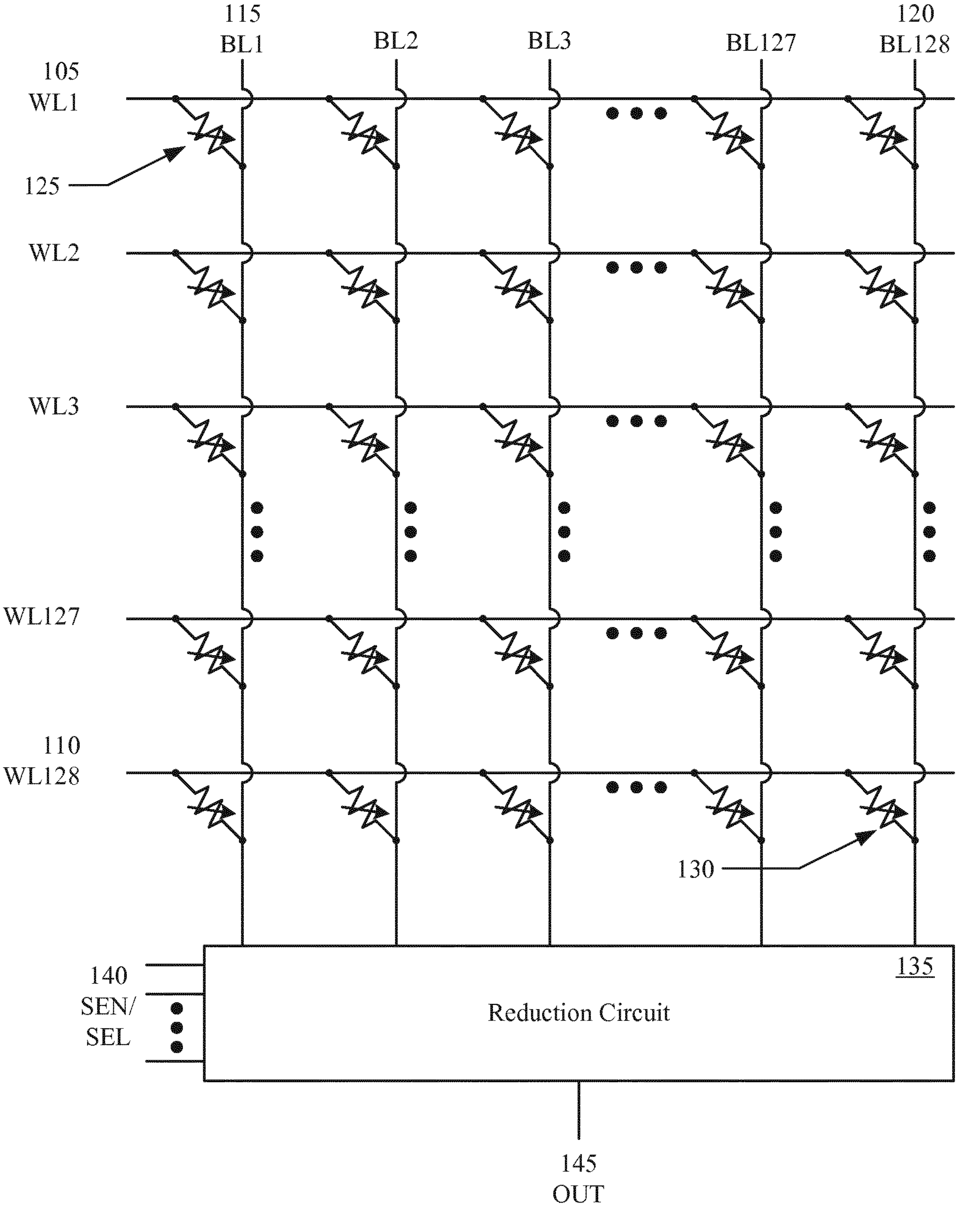
115
BL1
BL2
BL3
BL127
120
BL128
105
WL1
125
WL2
WL3
WL127
110
WL128
130
140
SEN/
SEL
135
Reduction Circuit
145
OUT

FIG. 1

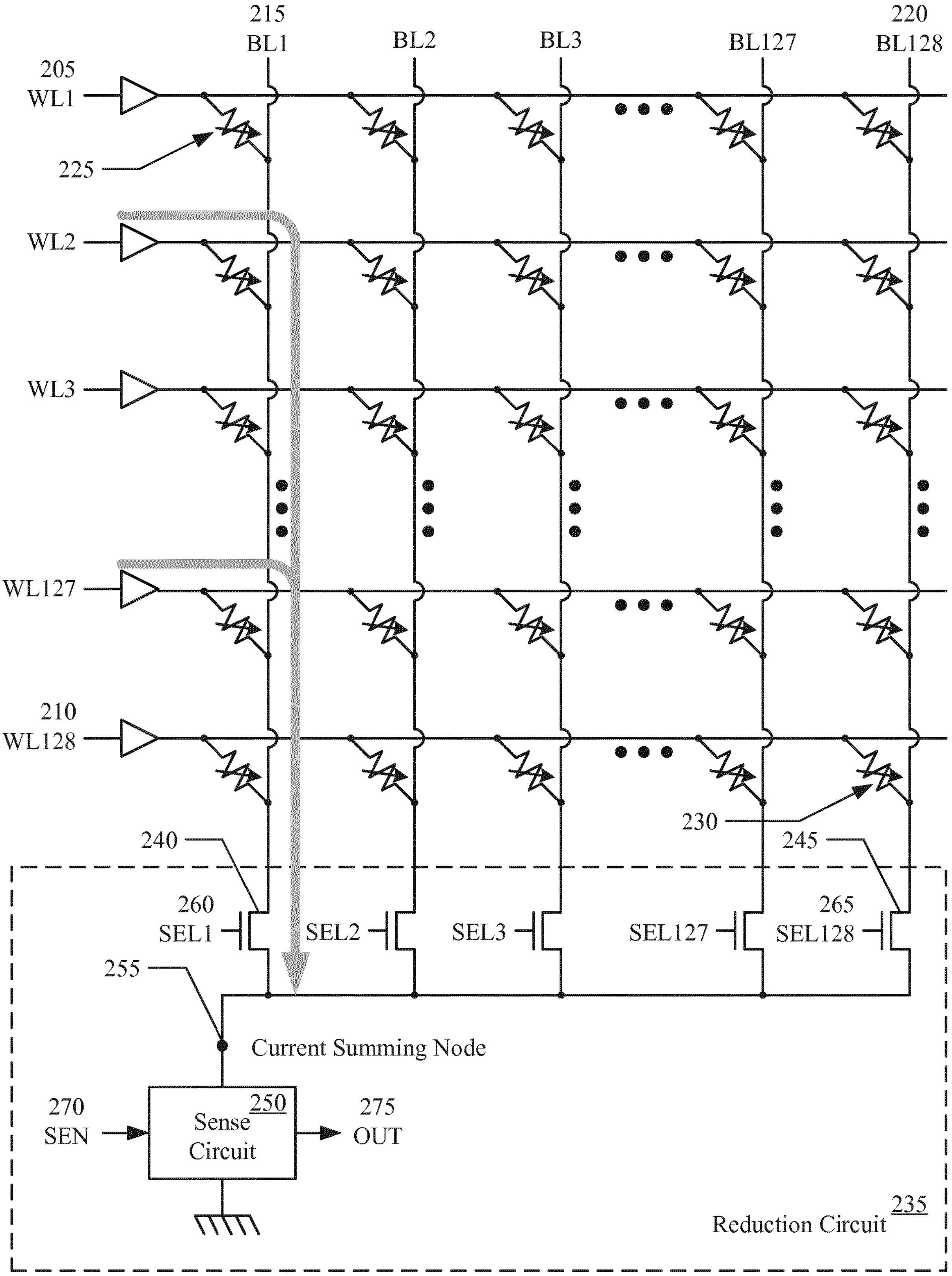
215
BL1
BL2
BL3
BL127
220
BL128
205
WL1
225
WL2
WL3
WL127
210
WL128
230
240
245
260
SEL1
SEL2
SEL3
SEL127
265
SEL128
255
Current Summing Node
270
SEN
250
Sense
Circuit
275
OUT
235
Reduction Circuit

FIG. 2

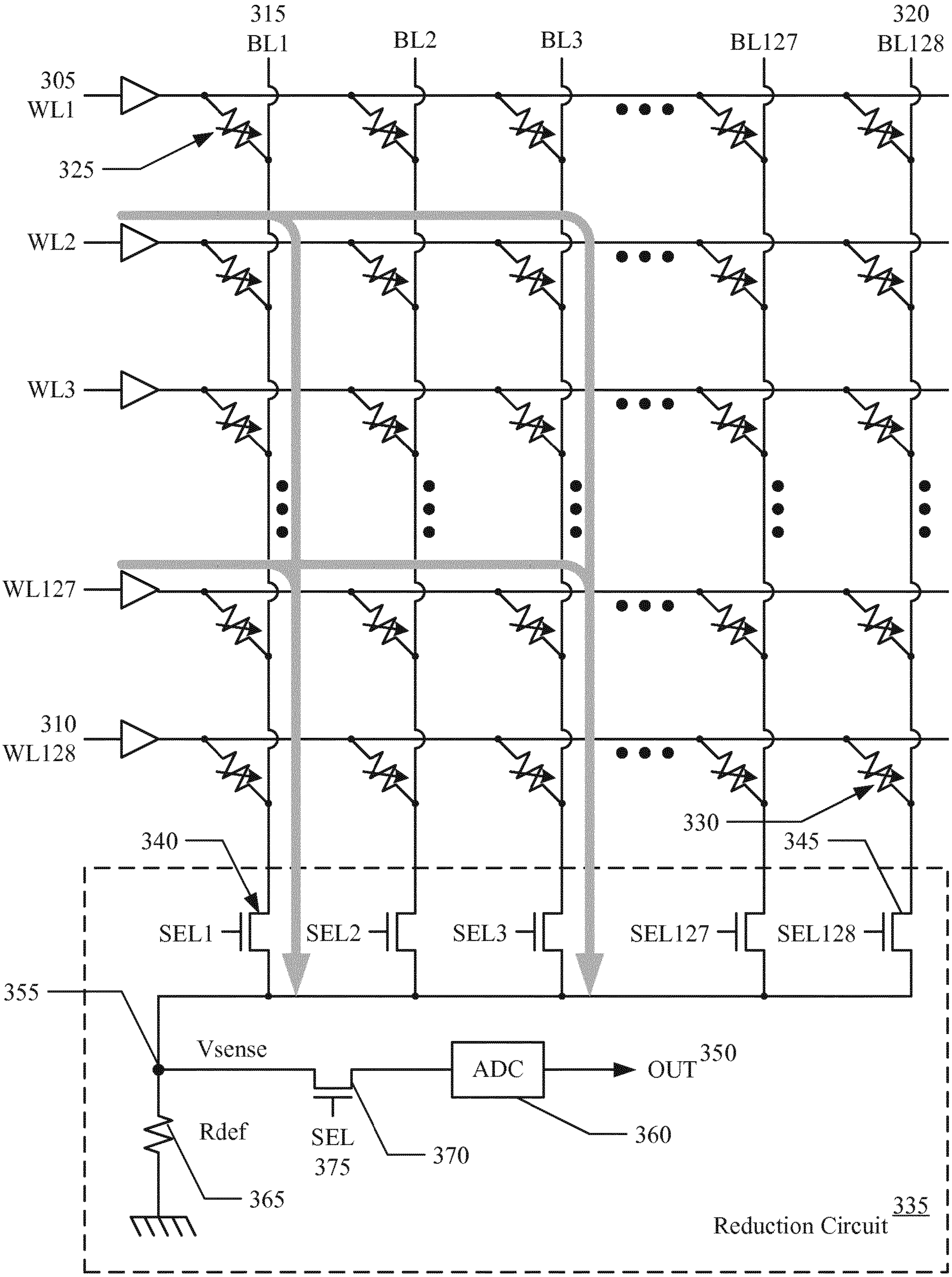
315
BL1
BL2
BL3
BL127
320
BL128
305
WL1
325
WL2
WL3
WL127
310
WL128
340
330
345
SEL1
SEL2
SEL3
SEL127
SEL128
355
Vsense
ADC
OUT
350
Rdef
SEL
375
370
360
365
335
Reduction Circuit

FIG. 3

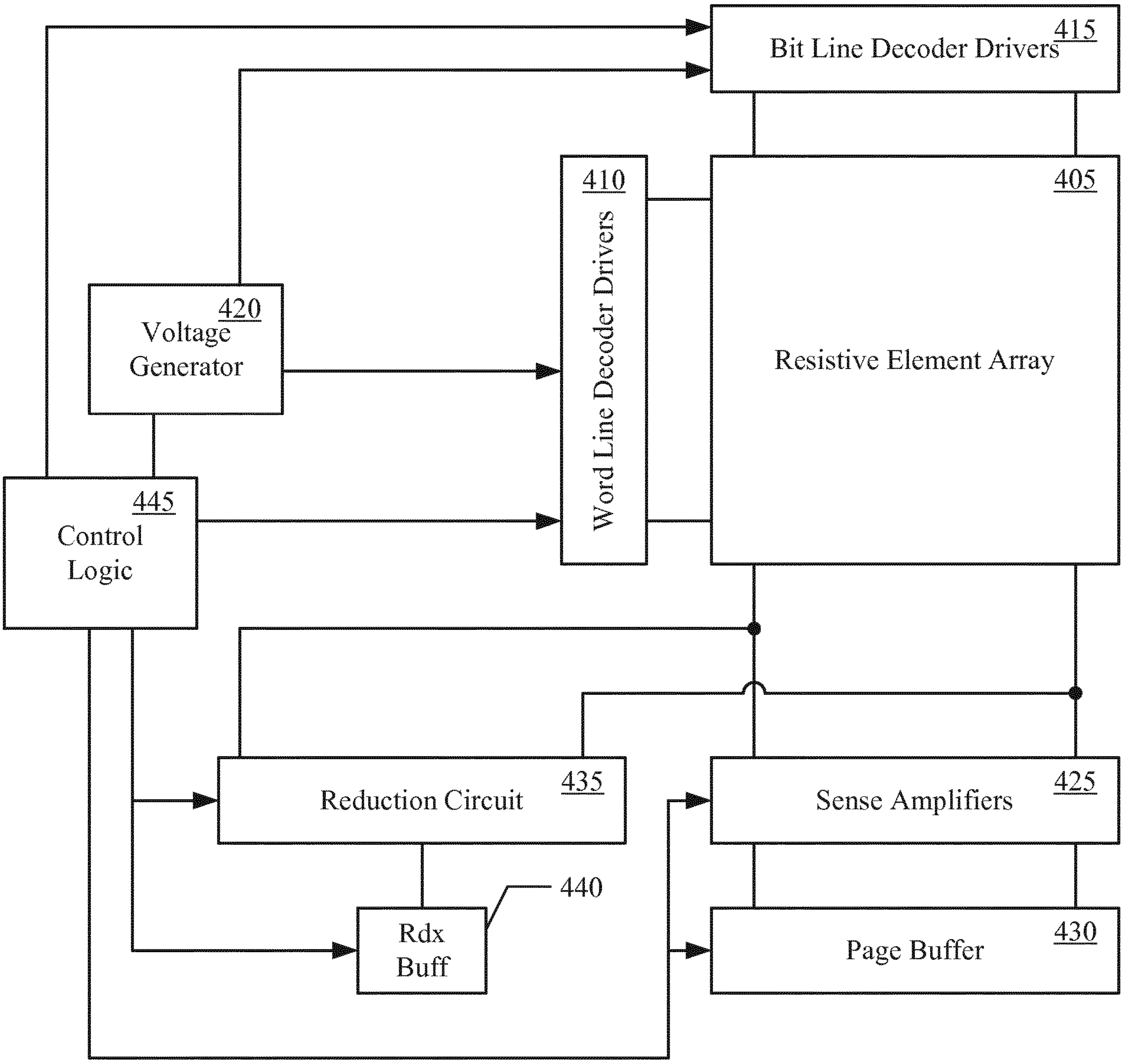
Bit Line Decoder Drivers
415
410
Word Line Decoder Drivers
405
Resistive Element Array
420
Voltage
Generator
445
Control
Logic
435
Reduction Circuit
425
Sense Amplifiers
440
Rdx
Buff
430
Page Buffer

FIG. 4

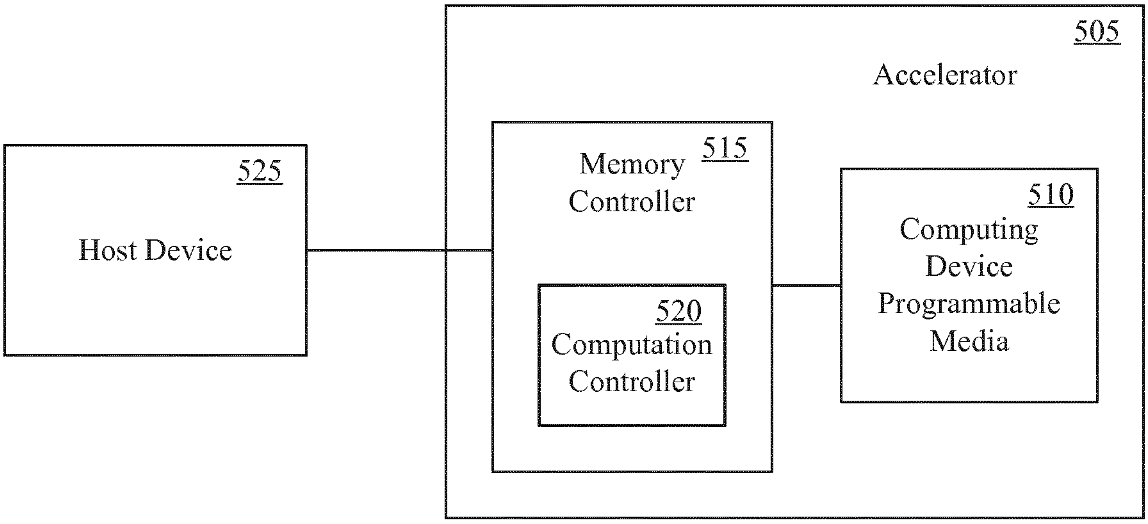
525
Host Device
505
Accelerator
515
Memory Controller
520
Computation Controller
510
Computing Device Programmable Media

FIG. 5

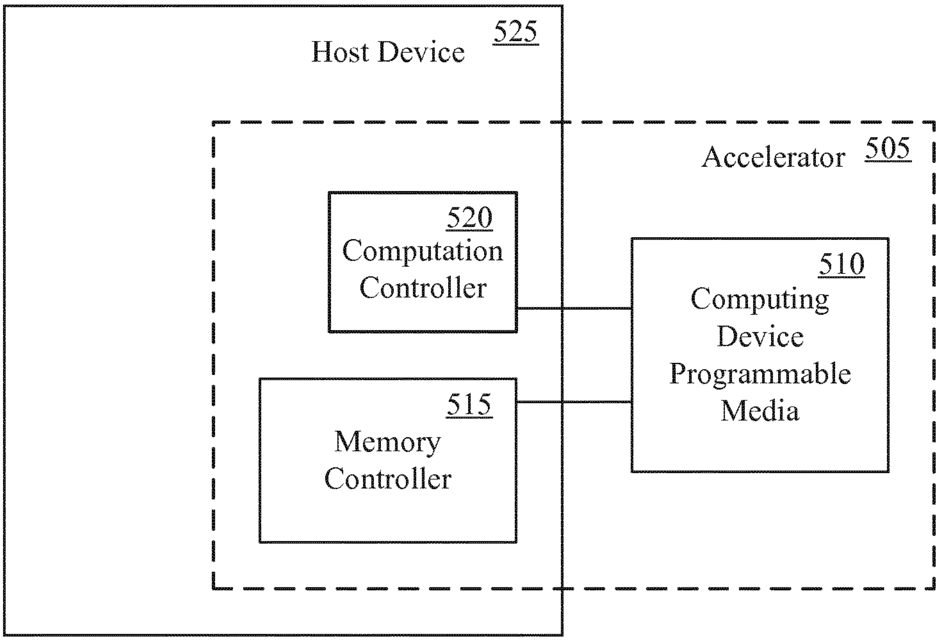
525
Host Device
Accelerator
505
520
Computation Controller
515
Memory Controller
510
Computing Device Programmable Media

FIG. 6

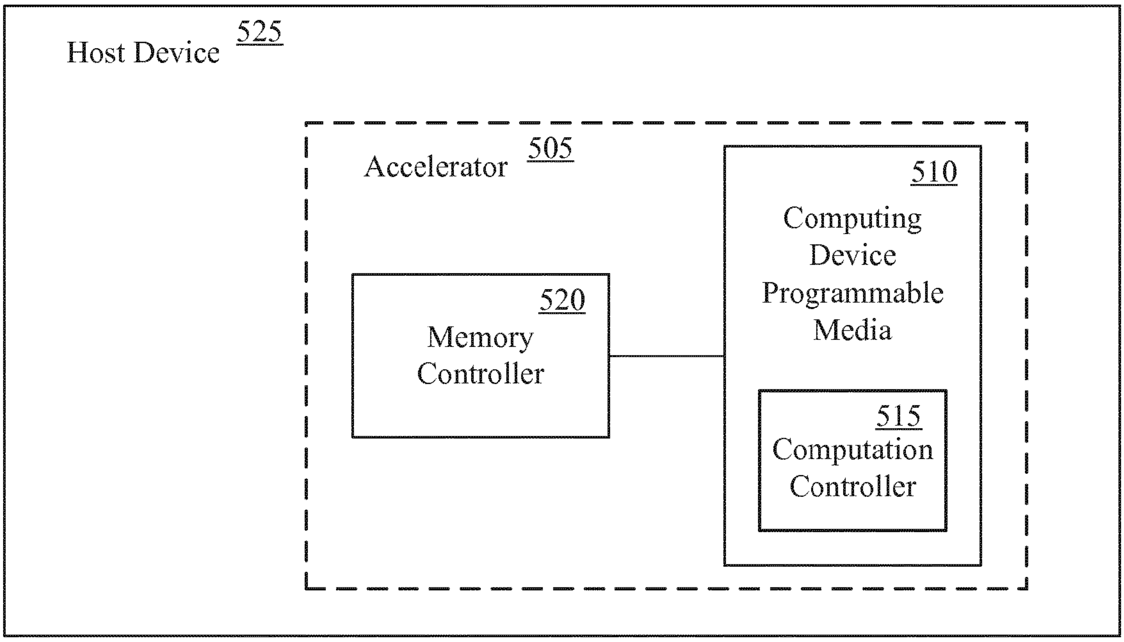
Host Device 525
Accelerator 505
520
Memory Controller
510
Computing Device Programmable Media
515
Computation Controller

FIG. 7

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- US 2016133321 A1 **[0002]**
- WO 2016099438 A1 **[0003]**
- WO 2017105517 A1 **[0006]**
- US 2017228345 A1 **[0007]**

Non-patent literature cited in the description

- Design considerations for variation tolerant multilevel CMOS/Nano memristor memory. **HARIKA MANEM et al.** GREAT LAKES SYMPOSIUM ON VLSI. ACM, 16 May 2010, 287-292 **[0004]**
- Energy-efficient SQL query exploiting RRAM-based process-in-memory structure. **SUN YULIANG et al.** IEEE 6TH NON-VOLATILE MEMORY SYSTEMS AND APPLICATIONS SYMPOSIUM (NVMSA). IEEE, 16 August 2017, 1-6 **[0005]**